Lời cảm ơn

Đầu tiên, em xin được gửi lời cảm ơn chân thành đến các thầy giáo, cô giáo thuộc trường đại học Bách Khoa Hà Nội, đặc biệt là các thầy giáo, cô giáo thuộc Viện Công nghệ Thông tin và Truyền Thông. Thầy cô là người đã trang bị cho em những kiến thức quý báu trong thời gian em học tập tại trường.

Em xin gửi lời cảm ơn sâu sắc tới thầy Ngô Văn Linh, người đã chỉ dẫn tận tình cho em trong suốt thời gian dài học tập tại trường. Thầy đã truyền đạt cho em rất nhiều kiến thức về topic modeling cũng như và nhiều chủ đề thú vị khác trong học máy. Thầy là người đã truyền cho em niềm hứng thú, sự đam mê tìm hiểu về học máy nói chung và topic modeling nói riêng từ đó dẫn đến sự hình thành của đồ án này. Thầy cũng là người trực tiếp hướng dẫn em trong quá trình thực hiện đồ án.

Em cũng xin gửi lời cảm ơn tới anh Mai Tiến Khải đã cũng cấp bộ dữ liệu quý giá giúp em có thể chạy thử nghiệm từ đó đánh giá phương pháp được đề xuất.

Tóm tắt đồ án

Ngày nay dữ liệu văn bản được sinh ra ngày càng nhiều thông qua các trang mạng xã hội, các diễn đàn và các dịch vụ chat, nhắn tin trực tuyến. Tuy nhiên, các dữ liệu này đều có chung một đặc điểm là lượng dữ liệu rất nhiều tuy nhiến độ dài của mỗi văn bản thường rất ngắn (gọi là short text). Điều này làm cho bài toán mô hình hóa chủ đề trở nên khó khăn hơn vì chi phí tính toán và đặc biệt là thiếu các thông tin thống kê về ngữ cảnh. Các phương pháp mô hình hóa chủ đề cổ điển như Latent Dirichlet Allocation (LDA) tuy có thể được thích ứng để làm việc với lượng dữ liệu lớn nhưng lại chịu ảnh hưởng của tính thưa của dữ liệu, điều này làm cho kết quả thu được từ LDA vẫn chưa thực sự đáp ứng được yêu cầu thực tế.

Một trong những phương pháp mô hình hóa chủ đề cho văn bản ngắn đang nhận được nhiều sự chú ý trong thời gian gần đầy là mô hình Biterm. Mô hình đã chứng minh được sự hiệu quả của mình khi làm việc với short text. Trong đồ án này, em sẽ phát triển dạng Online cho mô hình Biterm, cho phép mô hình có thể hoạt động với lượng dữ liệu lớn. Đánh giá về thời gian chạy và chất lượng chủ đề cũng sẽ được trình bày.

Mục lục

Danh sách các từ viết tắt và thuật ngữ

LDA	Latent Dirichlet Allocation
PLSA	Probabilistic Latent Semantic Analysis
BTM	Biterm Topic Model
Dir	Phân phối Dirichlet
Mult	Phân phối Multinomial
ML	Maximum Likelihood
MAP	Maximum A Posteriori
NPMI	Normalized Pointwise Mutual Information
LPBP	Trung bình log xác suất sinh ra một biterm
Corpus	Tập các văn bản
Doc	Văn bản
Topic	Chủ đề
Posterior distribution	phân phối hậu nghiệm
Prior distribution	Phân phối tiền nghiệm
Stochastic Optimization	Tối ưu hóa ngẫu nhiên
argmax	Giá trị tham số làm biểu thức lớn nhất
Parameter	Tham số
Hyperparameter	Siêu tham số
NYTT	Tập dữ liệu NewYorkTimes Title

Danh sách các kí hiệu dùng trong đồ án

Γ	hàm Gamma
$\sim$	"tuân theo phân phối"
$\propto$	"tỉ lệ với"

Danh sách hình vẽ

Danh sách bảng

1 Giới thiệu Đề Tài

2 Cơ sở lý thuyết

2.1 Mô hình chủ đề

2.1.1 Những khái niệm trong mô hình chủ đề

Khái niệm về chủ đề: Một chủ đề của dữ liệu có thể được hiểu theo định nghĩa thông thường, chẳng hạn như chủ đề về thể thao, chủ đề về chính trị, văn hóa, giáo dục ... Căn cứ vào đâu để biết một văn bản thuộc chủ đề nào ? Chính là căn cứ vào những từ xuất hiện trong văn bản mà ta có thể xác định chúng thuộc chủ đề gì. Nếu trong văn bản có chứa các từ như: *bóng đá, cầu thủ,trọng tài, thủ môn, hậu vệ ...* thì có thể nói là văn bản thuộc về chủ đề thể thao chứ không thể là thuộc về chủ đề về ẩm thực được. Như vậy có thể thầy bản chất của chủ đề là tập các từ mà chúng cùng xuất hiện với nhau một các thường xuyên trong một văn bản, mà khi nhìn vào văn bản đó ta thấy được sự nổi lên hẳn của các từ cùng xuất này và kết luận về chủ đề của nó. Trong mô hình chủ đề, cùng tư tưởng như vậy, một chủ đề cũng là tập các từ thường xuyên xuất hiện với nhau trong một văn bản nhưng được biểu diễn ở dạng toán học: một chủ đề được biểu diễn bằng một phân phối các từ trong tập từ điển, các từ khác nhau. Ví dụ như chủ đề về thể thao thì xác suất của từ *"bóng đá"* cao hơn xác suất của từ *"nhà hàng"* ngược lại thì chủ đề về ẩm thực thì xác suất của từ *"nhà hàng"* có xác suất lớn hơn so với từ *"bóng đá"*.

Vector tỉ lệ chủ để: Một bài báo thường có nhiều hơn một chủ đề, ví dụ như có những bài báo viết về chủ đề thể thao có liên quan đến chủ đề sức khỏe, thâm chí nó có các nội dung liên quan đến pháp luật chính trị. Giả sử như có K chủ đề, một văn bản sẽ là một phân phối trên K chủ đề n. Trong mô hình chủ đề, ta giả sử mỗi văn bản là tập trộn các chủ đề,

Tài liệu